Convergence Rate of K-Step Maximum Likelihood Estimate in Semiparametric Models

Guang Cheng

Duke University

Abstract: We suggest an iterative approach to computing K-step maximum likelihood estimates (MLE) of the parametric components in semiparametric models based on their profile likelihoods. The higher order convergence rate of K-step MLE mainly depends on the precision of its initial estimate and the convergence rate of the nuisance functional parameter in the semiparametric model. Moreover, we can show that the K-step MLE is as asymptotically efficient as the regular MLE after a finite number of iterative steps. Our theory is verified for several specific semiparametric models. Simulation studies are also presented to support these theoretical results.

Key words and phrases: K-Step Maximum Likelihood Estimate, Convergence Rate, Profile Likelihood, Semiparametric Models.

1. Introduction

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables from a semiparametric model $\mathbf{P} = \{P_{\theta,\eta} : \theta \in \Theta, \eta \in \mathcal{H}\}$, where θ is a $d-$dimensional parameter of interest and η is an infinite dimensional nuisance parameter. A well-known method of estimating the parameter θ in a semiparametric model is to solve θ from the below estimation equation:

$$\sum_{i=1}^{n} \tilde{\ell}_{\theta,\tilde{\eta}_n}(X_i) = 0, \tag{1.1}$$

where $\tilde{\eta}_n$ is some estimator for the nuisance parameter, and $\tilde{\ell}_{\theta,\eta}$ is the efficient score function for θ, whose definition will be introduced later. However, there are at least two concerns in solving (1.1). Firstly, we may have multiple roots in which identifying the consistent solution could be very challenging. Secondly, the above estimation approach requires an explicit form of the efficient score function, which in general is implicitly defined as an orthogonal projection. Although

we can estimate θ only by solving $\sum \dot{\ell}_{\theta,\eta_0}(X_i) = 0$, where $\dot{\ell}_{\theta,\eta_0}$ is the regular score function for θ given the true parameter η_0, in the semiparametric models of convex parametrization (page 305 in Bickel, Klaassen, Ritov and Wellner (1998)), many other semiparametric models of interest do not possess such nice properties.

The above concerns can be addressed well by the profile likelihood based K-step maximum likelihood estimate proposed in this paper. Under fairly general assumptions the K-step MLE is shown to posses higher order asymptotic efficiency than MLE of θ in semiparametric models. Actually the motivation for constructing k-step estimator $\hat{\theta}_n^{(k)}$ comes from the Newton-Raphson algorithm for solving (1.1) with respect to θ, starting at the initial guess $\hat{\theta}_n^{(0)}$. Thus, we can define k-step estimator iteratively in the below form:

$$\hat{\theta}_n^{(k)} = \hat{\theta}_n^{(k-1)} + \left(\mathbb{P}_n \tilde{\ell}_{\hat{\theta}_n^{(k-1)},\tilde{\eta}_n} \tilde{\ell}^T_{\hat{\theta}_n^{(k-1)},\tilde{\eta}_n}\right)^{-1} \mathbb{P}_n \tilde{\ell}_{\hat{\theta}_n^{(k-1)},\tilde{\eta}_n} \tag{1.2}$$

for $k = 1, 2, \ldots$, $\mathbb{P}_n f = \sum_{i=1}^n f(X_i)/n$ and $\hat{\theta}_n^{(0)}$ is some preliminary estimator for θ. In the parametric models, K-step MLE is defined similarly but with the efficient score function replaced by the regular score function for θ in (1.2). Under some regularity conditions in parametric models, Jassen, Jureckova and Veraverbeke (1985) shows that

$$\hat{\theta}_n^{(1)} - \hat{\theta}_n = O_P(n^{-1}) \text{ and } \hat{\theta}_n^{(2)} - \hat{\theta}_n = O_P(n^{-3/2}), \tag{1.3}$$

where $\hat{\theta}_n$ is maximum likelihood estimate for θ. The previous studies (Bickel, Klaassen, Ritov and Wellner (1998) and Van der Vaart (1998)) about K-step MLE only focus on the semiparametric models with convex parametrization, in which the efficient score functions can be estimated explicitly. Given certain no-bias conditions of the estimated efficient score functions, Van der Vaart (1998) shows that $\hat{\theta}_n^{(1)} = \hat{\theta}_n + o_P(n^{-1/2})$. Moreover, K-step approach is also used in local (quasi) likelihood estimation for the purpose of reducing computational cost, see Fan and Chen (1999), Fan, Chen and Zhou (2006) and Cai, Fan and Li (2000). However, as far as we are aware, it appears that no systematic studies have been done on the construction of K-step semiparametric MLE and its higher order asymptotic efficiency so far.

The efficient score function $\tilde{\ell}_{\theta,\eta}$ in (1.2) usually does not have an explicit

form or cannot be estimated explicitly as discussed above. Hence, we estimate $\mathbb{P}_n\tilde{\ell}_{\theta,\tilde{\eta}_n}$ and $\mathbb{P}_n\tilde{\ell}_{\theta,\tilde{\eta}_n}\tilde{\ell}^T_{\theta,\tilde{\eta}_n}$ via numerical derivatives of the profile likelihood. The profile likelihood $pl_n(\theta)$ is defined as $\sup_{\eta\in\mathcal{H}} lik_n(\theta,\eta)$, where $lik_n(\theta,\eta)$ is the full likelihood given n observations. In practice, the profile likelihood may have an explicit form, e.g. the Cox model with right censored data, or can be easily computed using procedures such as the fixed-point algorithm (as used in Kosorok, Lee and Fine (2004), for example) or the iterative convex minorant algorithm introduced in Groeneboom (1991) if η is a monotone function. Hence we will assume throughout this paper that evaluation of $pl_n(\theta)$ is computationally feasible. We shall consider the profile likelihood based K-step MLE in the form:

$$\hat{\theta}_n^{(k)} \;=\; \hat{\theta}_n^{(k-1)} + \Big(\Pi_n(\hat{\theta}_n^{(k-1)}, t_n)\Big)^{-1}\Gamma_n(\hat{\theta}_n^{(k-1)}, s_n) \tag{1.4}$$

for $k = 1, 2, \ldots$ and reasonably accurate starting point $\hat{\theta}_n^{(0)}$. $\Gamma_n(\theta, s_n)$ and $\Pi_n(\theta, t_n)$ are thus the discretized version of first and second derivative of the profile likelihood around θ with step size s_n and t_n, respectively. Their forms are given and justified in section 3. In section 2, we provide some necessary background about semiparametric models and two primary assumptions needed in this paper. In section 3, we discuss the construction of the initial estimates and present the main result of the paper about higher order convergence rate of K-step semiparametric MLE. In section 4, the proposed K-step approach is applied to three semiparametric models. Section 5 contains some simulations results of the Cox regression model, and proofs are given in section 6.

2. Background and Assumptions

We assume the data $X_1, \ldots, X_n$ are i.i.d. throughout the paper. In what follows, we first briefly review the concept of the efficient score function and define the convergence rate for the nuisance functional parameter. Next, we present two primary assumptions about second order asymptotic expansions of log-profile likelihood and MLE.

2.1 Preliminary

The *score function* for θ, $\dot{\ell}_{\theta,\eta}$, is defined as the partial derivative w.r.t. θ of the log-likelihood given η is fixed for a single observation. We denote the true

values of (θ, η) as (θ_0, η_0). A score function for η_0 is of the form

$$\frac{\partial}{\partial t}|_{t=0} \log p_{\theta_0,\eta_t}(x) \equiv A_{\theta_0,\eta_0} h(x),$$

where h is a "direction" by which $\eta_t \in \mathcal{H}$ approaches η_0, running through some index set H. $A_{\theta,\eta} : H \mapsto L_2^0(P_{\theta,\eta})$ is the score operator for η. The *efficient score function* for θ is defined as $\tilde{\ell}_{\theta,\eta} = \dot{\ell}_{\theta,\eta} - \Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta}$, where $\Pi_{\theta,\eta}\dot{\ell}_{\theta,\eta}$ minimizes the squared distance $P_{\theta,\eta}(\dot{\ell}_{\theta,\eta} - k)^2$ over all functions k in the closed linear space of the score functions for η (the "nuisance scores"). The inverse of the variance of $\tilde{\ell}_{\theta,\eta}$ is the Crámer Rao bound for estimating θ in the presence of the infinite dimensional nuisance parameter η, called efficient information matrix $\tilde{I}_{\theta,\eta}$. We also abbreviate $\tilde{\ell}_{\theta_0,\eta_0}$ and $\tilde{I}_{\theta_0,\eta_0}$ with $\tilde{\ell}_0$ and $\tilde{I}_0$, respectively. An insightful review of efficient score functions can be found in chapter 3 of Kosorok (2007).

The maximum likelihood estimate for (θ, η) can be expressed as $(\hat{\theta}_n, \hat{\eta}_n)$, where $\hat{\eta}_n = \hat{\eta}_{\hat{\theta}_n}$ and $\hat{\eta}_\theta = argmax_{\eta\in\mathcal{H}} lik_n(\theta, \eta)$. The convergence rate for η is defined as the largest r that satisfies $\|\hat{\eta}_{\tilde{\theta}_n} - \eta_0\| = O_P(\|\tilde{\theta}_n - \theta_0\| + n^{-r})$, where $\|\cdot\|$ is a norm with definition depending on context, i.e., for a Euclidean vector u, $\|u\|$ is the Euclidean norm, and for an element of the nuisance parameter space $\eta \in \mathcal{H}$, $\|\eta\|$ is some chosen norm on $\mathcal{H}$. In regular semiparametric models, which we can define without loss of generality to be models where the entropy integral converges, r is always larger than $1/4$. We say the nuisance parameter has parametric rate if $r = 1/2$. For instance, the nuisance parameters of the three examples in Cheng and Kosorok (2006) achieve the parametric rate. More specifically, the nuisance parameter in the Cox model, which is the cumulative hazard function, has the parametric rate under right censored data. However, the convergence rate for the cumulative hazard becomes slower, i.e. $r = 1/3$, under current status data.

2.2 Assumptions

The main result of this paper is based on the following second order asymptotic expansion of the profile likelihood, i.e. (2.1). For any random sequence

$\tilde{\theta}_n = \theta_0 + o_P(1)$, Cheng and Kosorok (2007) proves that

$$\begin{aligned} logpl_n(\tilde{\theta}_n) &= \log pl_n(\theta_0) + (\tilde{\theta}_n - \theta_0)^T \sum_{i=1}^{n} \tilde{\ell}_0(X_i) \\ &\quad -\frac{n}{2}(\tilde{\theta}_n - \theta_0)^T \tilde{I}_0(\tilde{\theta}_n - \theta_0) + O_P(g_r(\|\tilde{\theta}_n - \hat{\theta}_n\|)), \end{aligned} \tag{2.1}$$

where $g_r(w) \equiv (nw^3 \vee n^{1-2r}w \vee n^{-2r+1/2})1\{1/4 < r < 1/2\} + (nw^3 \vee n^{-1/2})1\{r \geq 1/2\}$, under certain second order no-bias conditions. Under similar conditions the maximum likelihood estimate is asymptotically normal, and has the asymptotic expansion:

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \tilde{I}_0^{-1} \tilde{\ell}_0(X_i) + O_P(n^{-1/2} + n^{-2r+1/2}), \tag{2.2}$$

where $\tilde{I}_0$ is assumed to be strictly positive definite. Expansions (2.1) and (2.2) are essentially second order versions of (1.4) and (1.5), which justify using a semi-parametric profile likelihood as an ordinary likelihood, in Murphy and Van der Vaart (2000). Under second order conditions specified in section 2.3 of Cheng and Kosorok (2007), (2.1) and (2.2) have been shown to hold in several semi-parametric models, e.g. Cox regression and partly linear model, in Cheng and Kosorok (2006) and Cheng and Kosorok (2007). Therefore, we assume (2.1) and (2.2) as two primary assumptions needed for the remainder of the paper.

3. Main Results

We first present two general approaches to searching for the preliminary estimates. And then we discuss how to construct the estimates for $\mathbb{P}_n\tilde{\ell}_{\theta,\tilde{\eta}_n}$ and $\mathbb{P}_n\tilde{\ell}_{\theta,\tilde{\eta}_n}\tilde{\ell}^T_{\theta,\tilde{\eta}_n}$ in (1.4) based on the profile likelihoods. Finally the convergence rate of K-step MLE is given. Such higher order convergence rate results are of interest particularly in small- or moderate-sized samples. The conditions (2.1) and (2.2) are assumed to hold in this section.

3.1 Initial Estimate

The start-up estimator is usually required to have reasonably good precision in the above K-step approach. In the parametric models, $\hat{\theta}_n^{(0)}$ is required to be $\sqrt{n}$ consistent such that one- and two-step MLE can achieve the convergence

rate as shown in (1.3). In our semiparametric model set-up, we need the initial estimate to be n^{ψ} consistent for $0 < \psi \leq 1/2$. The $\sqrt{n}$ consistent estimate in parametric models can be obtained through M-estimation theorem, i.e. theorem 5.21 in Van der Vaart (1998), or derived case by case in different examples. In the semiparametric models where the ad-hoc estimation methods for $\hat{\theta}_n^{(0)}$ are unavailable, we provide two general search strategies for $\hat{\theta}_n^{(0)}$: one is through some MCMC sampling procedure, called the profile sampler Lee, Kosorok and Fine (2005); another is through the deterministic or stochastic grid search over the profile likelihood function.

The profile sampler is the MCMC sampling from the posterior of the profile likelihood, and was proposed for the purpose of obtaining frequentist inference of θ Lee, Kosorok and Fine (2005). However, here we can use this convenient MCMC sampling procedure to yield $\sqrt{n}$-consistent $\hat{\theta}_n^{(0)}$ and consistent estimate for $\tilde{I}_0$. Specifically speaking, under the conditions (1.4), (1.5) in Murphy and Van der Vaart (2000) and mild conditions on the prior specified in theorem 1 of Lee, Kosorok and Fine (2005), Lee, Kosorok and Fine (2005) shows that

$$\tilde{E}_{\theta|\tilde{X}}(\theta) = \hat{\theta}_n + o_P(n^{-1/2}), \tag{3.1}$$

$$\hat{I}_n(PS) = \tilde{I}_0 + o_P(1), \tag{3.2}$$

where $\tilde{E}_{\theta|\tilde{X}}(\theta)$ and $\hat{I}_n(PS)$ are the sample mean and the inverse of the sample variance of the profile sampler, respectively.

Next, we provide an alternative grid search method to establish the n^{ψ} consistent start-up estimator when the above profile sampling procedure is unavailable or time consuming. When the dimension of θ is not large, we will conduct a deterministic search of objective function $Q_n(\theta)$, which is defined as $(logpl_n(\theta)/n)$, at regularly spaced grid over the whole compact parameter space Θ. We summarize this idea in the below theorem 1. Meanwhile, we need to assume the asymptotic uniqueness of $\hat{\theta}_n$:

$$Q_n(\tilde{\theta}_n) - Q_n(\hat{\theta}_n) = o_P(1) \quad \text{implies} \quad \tilde{\theta}_n - \theta_0 = o_P(1), \tag{3.3}$$

for any random sequence $\{\tilde{\theta}_n\} \in \Theta$.

THEOREM 1 *Let $\mathcal{D}_n$ be a set of points θ_i^D regularly spaced throughout Θ with*

cardinality larger than $cn^{d\psi}$ *for some* $c > 0$. *Suppose that the parameter space* Θ *be a compact subset of* $\mathbb{R}^d$ *and (3.3) holds, then we have for* $0 < \psi \leq 1/4$

$$\theta_n^D - \theta_0 = O_P(n^{-\psi}), \tag{3.4}$$

where $\theta_n^D = argmax_{\mathcal{D}_n} Q_n(\theta)$.

However, if the dimension d is very large, we prefer the outcome of a stochastic search whose search points are formed by the realizations of an independent random variable $\bar{\theta}$ with strictly positive density around θ_0.

COROLLARY 1 *Assume that* $\bar{\theta}$ *is independent of* $Q_n(\theta)$ *for all* $\theta \in \Theta$ *and admits a density having support* Θ *and bounded away from zero in some neighborhood of* θ_0. *Let* $\mathcal{S}_n$ *be a set of independent copies of* $\bar{\theta}$ *with cardinality larger than* $cn^{2\psi}$ *for some* $c > 0$. *Suppose that the parameter space* Θ *be a compact subset of* $\mathbb{R}^d$ *and (3.3) holds, then we have for* $0 < \psi \leq 1/4$

$$\theta_n^S - \theta_0 = O_P(n^{-\psi}), \tag{3.5}$$

where $\theta_n^S = argmax_{\mathcal{S}_n} Q_n(\theta)$.

3.2 K-step MLE

Before proceeding to give the convergence rate of K-step MLE, we first specify the forms of $\Gamma_n(\theta, s_n)$ and $\Pi(\theta, t_n)$ in (1.4). The intuitive idea behind the constructions of the estimators for $\mathbb{P}_n \tilde{\ell}_{\theta,\tilde{\eta}_n}$ and $\mathbb{P}_n \tilde{\ell}_{\theta,\tilde{\eta}_n} \tilde{\ell}^T_{\theta,\tilde{\eta}_n}$ is to use $\hat{\eta}_\theta$ as $\tilde{\eta}_n$ when making inferences about θ.

Specifically speaking, the ith component of $\Gamma_n(\theta, s_n)$ is constructed in the form:

$$\begin{aligned} [\Gamma_n(\theta, s_n)]_i &= \mathbb{P}_n \left\{ \frac{\log lik(\theta + s_n v_i, \hat{\eta}_{\theta + s_n v_i}) - \log lik(\theta, \hat{\eta}_\theta)}{s_n} \right\} \\ &= \frac{\log pl_n(\theta + s_n v_i) - \log pl_n(\theta)}{n s_n}, \end{aligned} \tag{3.6}$$

where step size $s_n \xrightarrow{P} 0$ and v_i denotes the ith unit vector in $\mathbb{R}^d$. Following similar

logic, we can define the (i,j)-th component of $\Pi_n(\theta, t_n)$ as:

$$[\Pi_n(\theta,t_n)]_{i,j} = - \frac{\log pl_n(\theta + v_i t_n + v_j t_n) + \log pl_n(\theta)}{nt_n^2} + \frac{\log pl_n(\theta + v_i t_n) + \log pl_n(\theta + v_j t_n)}{nt_n^2}, \tag{3.7}$$

where step size $t_n \xrightarrow{P} 0$. (3.7) is also called observed profile information in Murphy and Van der Vaart (1999). The lemma 1 in the appendix justifies the use of (3.6) and (3.7) as consistent estimates of $\mathbb{P}_n\tilde{\ell}_0$ and $\tilde{I}_0$, respectively.

The convergence rate of K-step MLE is certainly determined by the order of the step sizes in numerical differentiations $\Gamma_n(\cdot, s_n)$ and $\Pi_n(\cdot, t_n)$ as shown in the above. However, we are mostly interested in the fastest convergence rate K-step MLE can attain. Hence, we assume using the optimal step sizes (s_n^*, t_n^*), under which the fastest convergence rate of $\hat{\theta}_n^{(k)}$ is achieved, in the below theorem 2. As the theoretical basis for using K-step approach in practice, the below theorem 2 first presents the convergence rate for the fully iterative estimate $\hat{\theta}_n^{(\infty)}$, called optimal rate of K-step MLE, and then gives the number of iterations needed in (1.4) for $\hat{\theta}_n^{(k)}$ to attain the above optimal rate. Note that the asymptotic efficiency of $\hat{\theta}_n^{(k)}$ has continuously improved through the whole iterative procedure until it reaches the optimal bound based on the proof of theorem 2.

THEOREM 2 *Assume that $\hat{\theta}_n^{(k)}$ is defined as (1.4) and $\hat{\theta}_n^{(0)}$ is n^{ψ}-consistent for $0 < \psi \leq 1/2$, we have*

$$\hat{\theta}_n^{(\infty)} - \hat{\theta}_n = O_P(n^{-3/4} \vee n^{-r-1/4}). \tag{3.8}$$

Moreover, the above optimal rate can be achieved after N (M) iterations starting from $\hat{\theta}_n^{(0)}$ in (1.4) for $r \geq 1/2$ $(1/4 < r < 1/2)$:

$$\hat{\theta}_n^{(N)} - \hat{\theta}_n = O_P(n^{-3/4}), \tag{3.9}$$

$$\hat{\theta}_n^{(M)} - \hat{\theta}_n = O_P(n^{-r-1/4}), \tag{3.10}$$

where $N = int[\log 2\psi/\log(2/3)]+1$, $M = int[\log(\psi/r)/\log(2/3)]+int[\log(4r/(4r-1))/\log(2) - 1] + 1$ and $int[x]$ indicates the smallest nonnegative integer larger than or equal to x.

From theorem 1 and 2, it is not surprising to find that there exists a tradeoff between the number of search grids and the number of iterations. Combining (3.9) and (3.10) with (2.2), we have the following asymptotic expansion of K-step MLE:

$$\begin{aligned}\sqrt{n}(\hat{\theta}_n^{(N)} - \theta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{I}_0^{-1}\tilde{\ell}_0(X_i) + O_P(n^{-1/4}),\\ \sqrt{n}(\hat{\theta}_n^{(M)} - \theta_0) &= \frac{1}{\sqrt{n}}\sum_{i=1}^{n} \tilde{I}_0^{-1}\tilde{\ell}_0(X_i) + O_P(n^{-r+1/4}).\end{aligned}$$

Thus we can construct the $(1-\alpha)$-th two sided asymptotically correct confidence interval for θ based on K-step MLE, i.e. $(\hat{\theta}_n^{(k)} - z_{1-\alpha/2}/\sqrt{n\tilde{I}}, \hat{\theta}_n^{(k)} + z_{1-\alpha/2}/\sqrt{n\tilde{I}})$, where $k = M$ or N, z_α is the standard normal α-th quantile, and $\tilde{I}$ is a consistent estimator of $\tilde{I}_0$.

REMARK 1 *Recall that in the parametric models, Jassen, Jureckova and Veraverbeke (1985) shows that* $\hat{\theta}_n^{(1)} - \hat{\theta}_n = O_P(n^{-1})$ *and* $\hat{\theta}_n^{(2)} - \hat{\theta}_n = O_P(n^{-3/2})$. *However, the optimal rate for the K-step MLE is slower even in the semiparametric models with parametric convergence rate. Such efficiency loss can be partially explained by the less smoothness of the profile likelihood in semiparametric models. In other words, the corresponding estimators for the score function and information matrix in K-step parametric MLE usually have bias of smaller order.*

4. Examples

In this section, the above K-step estimation approach is illustrated with three semiparametric models of different convergence rates. Under the model assumptions specified in section 5 of Cheng and Kosorok (2007), Cheng and Kosorok (2007) shows that (2.1) and (2.2) hold in all the examples. Hence, we only briefly review the model set-up here, and then discuss the choices of the initial estimates. Finally, we apply the theorem 2 to figure out the least number of iterations in K-step MLE needed to achieve the full efficiency.

4.1 Cox regression with right censored data

In the Cox regression model, the hazard function of the survival time T of a

subject with covariate Z is expressed as:

$$\lambda(t|z) \equiv \lim_{\Delta \to 0} \frac{1}{\Delta} Pr(t \leq T < t + \Delta | T \geq t, Z = z) = \lambda(t) \exp(\theta z), \tag{4.1}$$

where λ is an unspecified baseline hazard function and θ is a vector including the regression parameters Cox (1972). Under right censoring data, we only know that the event time T has occurred either before the censoring time C, or after the censoring time C. More precisely, the data observed is $X = (Y, \delta, Z)$, where $Y = T \wedge C$, $\delta = I\{T \leq C\}$, and $Z \in \mathbb{Z} \subset \mathbb{R}$ is a regression covariate. In the Cox regression model, we are usually interested in the regression parameter θ while treating the cumulative hazard function η as the nuisance parameter. Thus we express the likelihood for (θ, η) in the below form:

$$lik(\theta, \eta) = \left(e^{\theta z} \eta\{y\} e^{-e^{\theta z} \eta(y)} \right)^{\delta} \left(e^{-e^{\theta z} \eta(y)} \right)^{1-\delta}, \tag{4.2}$$

by replacing hazard function $\lambda(y)$ by the point mass $\eta\{y\}$. By the special construction of the Cox model, we have an explicit form of the log-profile likelihood:

$$\log pl_n(\theta) = \sum_{i=1}^{l} (\theta z_{[i]} - \log \sum_{j \in R_i} e^{\theta z_j}), \tag{4.3}$$

where $R_i = \{j : Y_j \geq t_i\}$, t_i is the observed value of the i-th ordered event time and $z_{[i]}$ is the covariate corresponding to t_i. The convergence rate of the estimated nuisance parameter is established in theorem 3.1 of Murphy and Van der Vaart (1999):

$$\|\hat{\eta}_{\tilde{\theta}_n} - \eta_0\|_{\infty} = O_P(n^{-\frac{1}{2}} + \|\tilde{\theta}_n - \theta_0\|), \tag{4.4}$$

where $\|\cdot\|_{\infty}$ denotes the uniform norm.

In this model, the profile sampler is generated very fast because of the explicit form for the profile likelihood. Hence, we use it to yield the root-n consistent start-up estimator. By theorem 2, we can conclude that $\hat{\theta}_n^{(1)} - \hat{\theta}_n = O_P(n^{-3/4})$, where $\hat{\theta}_n^{(1)}$ is constructed according to (1.4).

4.2 Cox regression for current status data

Current status data arises when each subject is observed at a single exam-

ination time, Y, to determine if an event has occurred. The event time, T, cannot be known exactly. Then the observed data are n i.i.d. realizations of $X = (Y, \delta, Z) \in R^+ \times \{0, 1\} \times R$, where $\delta = I\{T \leq Y\}$ and Z is a vector of covariates. It is not difficult to derive the log-likelihood:

$$\log lik_n(\theta, \eta) = \sum_{i=1}^{n} \delta_i \log[1 - \exp(-\eta(Y_i) \exp(\theta Z_i))] - (1 - \delta_i) \exp(\theta Z_i)\eta(Y_i) \quad (4.5)$$

Moreover, using entropy methods, Murphy and Van der Vaart (1999) extends earlier results of Huang (1996), show that

$$\|\hat{\eta}_{\tilde{\theta}_n} - \eta_0\|_{L_2} = O_P(\|\tilde{\theta}_n - \theta_0\| + n^{-1/3}), \quad (4.6)$$

where $\|\cdot\|_{L_2}$ is the L_2 norm w.r.t. the distribution of Y.

In the Cox regression with current status data, the iterative convex minorant algorithm Huang (1996) is implemented to yield the profile likelihood. The MCMC sampling procedure thus becomes more time consuming because of such iterative computation mechanism. Hence, we prefer using grid search approach to obtain $n^{1/4}$-consistent preliminary estimate. We know that three step MLE attains the optimal rate, i.e. $\hat{\theta}_n^{(3)} - \hat{\theta}_n = O_P(n^{-7/12})$, based on theorem 2.

4.3 The partly linear model

In this model, a continuous outcome Y, conditional on the covariates $(W, Z) \in \mathbb{R}^d \times \mathbb{R}$, is modeled as:

$$Y = \theta^T W + k(Z) + \xi, \quad (4.7)$$

where k is an unknown smooth function, and $\xi \sim N(0, 1)$. The functional nuisance parameter k is assumed to belong to $\mathcal{O}_2 \equiv \{f : J_2(f) + \|f\|_\infty < M, \text{ for a known } M < \infty\}$, where $J_2(f)$ is the second order Sobolev norm of f. However, the response Y is not observed directly, but only its current status is observed at a random censoring time $C \in \mathbb{R}$. In other words, we observe $X = (C, \Delta, W, Z)$, where $\Delta = 1_{\{Y \leq C\}}$. Additionally (Y, C) is assumed to be independent given (W, Z). Under the model (4.7), the log-likelihood for a single

observation at $X = x \equiv (c, \delta, w, z)$ can be shown to have the form:

$$loglik_{\theta,k}(x) = \delta \log \{\Phi(c - \theta w - k(z))\} + (1 - \delta) \log \{1 - \Phi(c - \theta w - k(z))\} \quad (4.8)$$

where Φ is the standard normal distribution. In lemma 4 of Cheng and Kosorok (2007), we have shown

$$\left\| \hat{k}_{\tilde{\theta}_n} - k_0 \right\|_{L_2} = O_P(n^{-2/5} + \|\tilde{\theta}_n - \theta_0\|). \quad (4.9)$$

The rate $r = 2/5$ is clearly faster than the cubic rate but slower than the parametric rate. Depending on the dimension of θ, we can choose the deterministic or random search for the starting estimate. Similarly, we can show that four iterations are needed to achieve the optimal rate, i.e. $\hat{\theta}_n^{(4)} - \hat{\theta}_n = O_P(n^{-13/20})$, if $\hat{\theta}_n^{(0)}$ is $n^{1/4}$-consistent.

5. Simulations

It is of interest to see, at a finite sample, how good the K-step MLE is in comparison with the regular MLE. Hence, we conducted simulations in the Cox regression model with right censored data and with current status data in this section. The simulation results presented in the table 1 and 2 agree with our theoretical results given in subsection and .

We first run the simulations for various sample sizes in the Cox model with right censored data. As indicated in subsection , we can construct $\hat{\theta}_n^{(1)}$ in the form of (1.4) with (s_n^*, t_n^*) set to be proportional to $(n^{-3/4}, n^{-1/2})$ according to the proof of theorem 2. The profile sampler is generated under a Lebesgue prior. For each sample size, 500 datasets were analyzed. The event times were generated from (4.1) with one covariate $Z \sim U[0, 1]$. The regression coefficient is $\theta = 1$ and $\eta(t) = \exp(t) - 1$. The censoring time $C \sim U[0, t_n]$, where t_n was chosen such that the average effective sample size over 500 samples is approximately $0.9n$. For each dataset, Markov chains of length 5,000 with a burn-in period of 1,000 were generated using the Metropolis algorithm. The jumping density for the coefficient was normal with current iteration and variance tuned to yield an acceptance rate of $20\% - 40\%$. In the Cox regression with current status data, we first use the deterministic search over $[-5, 5]$ for the $n^{1/4}$ consistent $\hat{\theta}_n^{(0)}$. The three step MLE is iteratively generated according to (1.4), in which the order of

(s_n^*, t_n^*) at each step is specified in the proof of theorem 2.

In the appendix, the table 1 (2) summarizes the results from the simulations of Cox regression with right censored data (current status data) giving the average across 500 samples of K-step MLE and the maximum likelihood estimate (MLE). According to theorem 2, $n^{3/4}|\hat{\theta}_n - \hat{\theta}_n^{(1)}|$ ($n^{7/12}|\hat{\theta}_n - \hat{\theta}_n^{(3)}|$) in Cox model with right censored data (current status data) is bounded in probability. And the realizations of these terms summarized in table 1 and 2 clearly illustrate their boundedness. For each sample size, we can clearly observe that K-step MLE approaches to $\hat{\theta}_n$ after every iteration. Hence we can conclude that the numerical evidence in this section supports our theoretical results.

Acknowledgment The author thank Dr. Michael Kosorok for several insightful discussions.

6. Appendix

In the below lemma 1, we first provide a key technical tool for deriving higher order convergence rate of K-step MLE. The symbol $R_n \asymp q_n$ means that some random quantity $R_n = O_P(q_n)$ and $R_n^{-1} = O_P(q_n^{-1})$, where $q_n \to 0$.

Lemma 1. Assume the conditions (2.1) and (2.2) and $\hat{\theta}_n^{(0)}$ is a n^{ψ}-consistent estimate for $0 < \psi \leq 1/2$, then we have

$$\begin{aligned}
\Gamma_n(\hat{\theta}_n^{(0)}, s_n) &= \mathbb{P}_n\tilde{\ell}_0 + O_P\left(n^{-\psi} \vee |s_n| \vee \frac{g_r(n^{-\psi} \vee |s_n|)}{n|s_n|}\right), \quad (6.1)\\
\Gamma_n(\hat{\theta}_n^{(0)} + U_n, s_n) &= \Gamma_n(\hat{\theta}_n^{(0)}, s_n) - \tilde{I}_0 U_n \\
&\quad + O_P\left(|s_n| \vee \frac{g_r(n^{-\psi} \vee |s_n| \vee \|U_n\|)}{n|s_n|}\right), \quad (6.2)\\
\Pi_n(\tilde{\theta}_n, t_n) &= \tilde{I}_0 + O_P\left(\frac{g_r(r_n \vee |t_n|)}{nt_n^2}\right), \quad (6.3)
\end{aligned}$$

where $U_n = O_P(n^{-s})$ for some $s > 0$, $(\tilde{\theta}_n - \hat{\theta}_n) = O_P(r_n)$ and $g_r(w) \equiv (nw^3 \vee n^{1-2r}w \vee n^{-2r+1/2})1\{1/4 < r < 1/2\} + (nw^3 \vee n^{-1/2})1\{r \geq 1/2\}$.

Proof of lemma 1: (2.1) implies that

$$\begin{aligned} logpl_n(\hat{\theta}_n^{(0)} + V_n + s_n v_i) &= \log pl_n(\theta_0) + (\hat{\theta}_n^{(0)} + V_n + s_n v_i - \theta_0)^T \sum_{i=1}^{n} \tilde{\ell}_0(X_i) \\ &\quad -\frac{n}{2}(\hat{\theta}_n^{(0)} + V_n + s_n v_i - \theta_0)^T \tilde{I}_0(\hat{\theta}_n^{(0)} + V_n + s_n v_i - \theta_0) \\ &\quad +O_P(g_r(n^{-\psi} \vee |s_n| \vee \|V_n\|)), \\ logpl_n(\hat{\theta}_n^{(0)} + V_n) &= \log pl_n(\theta_0) + (\hat{\theta}_n^{(0)} + V_n - \theta_0)^T \sum_{i=1}^{n} \tilde{\ell}_0(X_i) \\ &\quad -\frac{n}{2}(\hat{\theta}_n^{(0)} + V_n - \theta_0)^T \tilde{I}_0(\hat{\theta}_n^{(0)} + V_n - \theta_0) \\ &\quad +O_P(g_r(n^{-\psi} \vee \|V_n\|)), \end{aligned}$$

for any random vector $V_n = o_P(1)$ and $s_n \xrightarrow{P} 0$. Combining the above two expansions and (3.6), we have

$$\Gamma_n(\hat{\theta}_n^{(0)} + V_n, s_n) = \mathbb{P}_n\tilde{\ell}_0 - \tilde{I}_0(\hat{\theta}_n^{(0)} - \theta_0) - \tilde{I}_0 V_n + O_P\left(|s_n| \vee \frac{g_r(n^{-\psi} \vee |s_n| \vee \|V_n\|)}{n|s_n|}\right).$$

By replacing $V_n = 0$ and $V_n = U_n$ in the above equation, we have proved (6.1) and (6.2), respectively. Taking into account (2.1) and (2.2), we can prove the below second order asymptotic expansion of the profile likelihood around $\hat{\theta}_n$:

$$\log pl_n(\tilde{\theta}_n) = \log pl_n(\hat{\theta}_n) - \frac{1}{2}n(\tilde{\theta}_n - \hat{\theta}_n)^T \tilde{I}_0(\tilde{\theta}_n - \hat{\theta}_n) + O_P(g_r(\|\tilde{\theta}_n - \hat{\theta}_n\|)) \quad (6.4)$$

for any sequence $\tilde{\theta}_n = \hat{\theta}_n + o_P(1)$. Following similar analysis in the above, (3.7) and (6.4) yield (6.3). This completes the whole proof. □

Proof of theorem 1: (2.1) implies that for $\tilde{\theta}_n - \theta_0 = o_P(1)$

$$Q_n(\tilde{\theta}_n) = Q_n(\theta_0) + (\tilde{\theta}_n - \theta_0)^T \mathbb{P}_n\tilde{\ell}_0 - \frac{1}{2}(\tilde{\theta}_n - \theta_0)^T \tilde{I}_0(\tilde{\theta}_n - \theta_0) + \Delta_n, \quad (6.5)$$

where $\Delta_n = O_P(g_r(\|\tilde{\theta}_n - \hat{\theta}_n\|))/n$. We then show that $P(\|\theta_n^D - \theta_0\| > Cn^{-\psi}) \to 0$ by the below set of inequalities for some $C > 0$. Set $\mathcal{N}_n = \{\theta : \|\theta - \theta_0\| \le Cn^{-\psi}\}$ and $\mathcal{N}_n^c$ denotes its complement. Note that $\mathcal{D}_n \cap \mathcal{N}_n \neq \emptyset$ for C large enough and

$\mathcal{D}_n \cap \mathcal{N}_n^c \neq \emptyset$ for n large enough.

$$\begin{aligned}
P(\theta_n^D \in \mathcal{N}_n^c) &\leq P\left(\max_{\mathcal{D}_n\cap\mathcal{N}_n} Q_n(\theta) \leq \max_{\mathcal{D}_n\cap\mathcal{N}_n^c} Q_n(\theta)\right) \\
&\leq P\left(\max_{\mathcal{D}_n\cap\mathcal{N}_n} Q_n(\theta) < Q_n(\theta_0) - C_1 n^{-2\psi}\right) \\
&\quad + P\left(\left\{\max_{\mathcal{D}_n\cap\mathcal{N}_n} Q_n(\theta) \leq \max_{\mathcal{D}_n\cap\mathcal{N}_n^c} Q_n(\theta)\right\} \cap \right. \\
&\qquad\qquad \left. \left\{\max_{\mathcal{D}_n\cap\mathcal{N}_n} Q_n(\theta) \geq Q_n(\theta_0) - C_1 n^{-2\psi}\right\}\right) \\
&\leq P\left(\max_{\mathcal{D}_n\cap\mathcal{N}_n} \sqrt{n}(Q_n(\theta) - Q_n(\theta_0)) < -C_1 n^{1/2-2\psi}\right) \\
&\quad + P\left(\max_{\mathcal{D}_n\cap\mathcal{N}_n^c} \sqrt{n}(Q_n(\theta) - Q_n(\theta_0)) \geq -C_1 n^{1/2-2\psi}\right),
\end{aligned}$$

where C_1 is some positive number. The first inequality in the above follows from the definition of θ_n^D. Based on (6.5) we have

$$\begin{aligned}
&P\left(\max_{\mathcal{D}_n\cap\mathcal{N}_n} \sqrt{n}(Q_n(\theta) - Q_n(\theta_0)) < -C_1 n^{1/2-2\psi}\right) \\
&= P\left(\max_{\mathcal{D}_n\cap\mathcal{N}_n} \left(\sqrt{n}(\theta-\theta_0)^T \mathbb{P}_n \tilde{\ell}_0 - \frac{\sqrt{n}}{2}(\theta-\theta_0)^T \tilde{I}_0 (\theta-\theta_0) + \sqrt{n}\Delta_n\right) < -C_1 n^{1/2-2\psi}\right) \\
&\leq P\left(\max_{\mathcal{D}_n\cap\mathcal{N}_n} (\theta-\theta_0)(-\sqrt{n}\mathbb{P}_n\tilde{\ell}_0) + \max_{\mathcal{D}_n\cap\mathcal{N}_n} ((\sqrt{n}/2)(\theta-\theta_0)^T \tilde{I}_0(\theta-\theta_0)) + \right. \\
&\qquad\qquad \left. \max_{\mathcal{D}_n\cap\mathcal{N}_n} (\sqrt{n}\Delta_n) > C_1 n^{1/2-2\psi}\right) \\
&\leq P\left(\sqrt{n}\mathbb{P}_n\tilde{\ell}_0 \gtrsim (C_1 - \delta C^2/2) n^{1/2-2\psi} + O_P(n^{1/2-3\psi})\right),
\end{aligned}$$

where δ is the largest eigenvalue for $\tilde{I}_0$. The last inequality in the above follows from the compactness of Θ. Let $\theta_n^* = argmax_{\mathcal{D}_n\cap\mathcal{N}_n^c} \sqrt{n} Q_n(\theta)$. (3.3) implies that

$\theta_n^* - \theta_0 = o_P(1)$ since $Q_n(\hat{\theta}_n) - Q_n(\theta_0) = o_P(1)$. Thus, by (6.5), we have

$$\begin{aligned}
&P\left(\max_{\mathcal{D}_n \cap \mathcal{N}_n^c} \sqrt{n}(Q_n(\theta) - Q_n(\theta_0)) \geq -C_1 n^{1/2-2\psi}\right) \\
&= P\left(\sqrt{n}(\theta_n^* - \theta_0)\mathbb{P}_n \tilde{\ell}_0 - \frac{\sqrt{n}}{2}(\theta_n^* - \theta_0)^T \tilde{I}_0 (\theta_n^* - \theta_0) + O_P(g_r(\|\theta_n^* - \hat{\theta}_n\|)/\sqrt{n})\right. \\
&\qquad\qquad \left. \geq -C_1 n^{1/2-2\psi}\right) \\
&\leq P\left(\sqrt{n}\mathbb{P}_n \tilde{\ell}_0 \gtrsim (\delta K_1^2/2 - C_1) n^{1/2-2\psi} + O_P(n^{1/2-3\psi})\right)
\end{aligned}$$

Note that θ_n^* belongs to the regularly spaced grid set $\mathcal{D}_n$ and $\|\theta_n^* - \theta_0\| > Cn^{-\psi}$. Therefore, we can conclude that θ_n^* should be the closest grid point to θ_0 but not in $\mathcal{N}_n$, i.e. $K_1 n^{-\psi} \leq \|\theta_n^* - \theta_0\| \leq K_2 n^{-\psi}$, where $C < K_1 \leq K_2 \leq 2C$ for large C, from (6.5) and the construction of $\mathcal{D}_n$. Without loss of generality, we assume $\theta_n^* > \theta_0$. Thus the last inequality in the above follows. Note that $\sqrt{n}\mathbb{P}_n \tilde{\ell}_0 = O_P(1)$ and $\psi \leq 1/4$. By choosing sufficiently large C and C_1, meanwhile keeping the inequality $\delta C^2/2 < C_1 < \delta K_1^2/2$ hold, we can $P(\theta_n^D \in \mathcal{N}_n^c) \to 0$ based on the above inequalities. □

Proof of corollary 1: The proof is similar to that of theorem 1. We still need to show that $P(\|\theta_n^S - \theta_0\| > Cn^{-\psi}) \to 0$ for some $C > 0$. Similarly, we have

$$\begin{aligned}
P(\theta_n^S \in \mathcal{N}_n^c) &\leq E\left[P\left(\max_{\mathcal{S}_n \cap \mathcal{N}_n} Q_n(\theta) \leq \max_{\mathcal{S}_n \cap \mathcal{N}_n^c} Q_n(\theta) | \mathcal{S}_n\right)\right] \\
&\leq E\left[P\left(\max_{\mathcal{S}_n \cap \mathcal{N}_n} \sqrt{n}(Q_n(\theta) - Q_n(\theta_0)) < -C_1 n^{1/2-2\psi} | \mathcal{S}_n\right)\right] \\
&\quad + P\left(\sup_{\mathcal{N}_n^c} \sqrt{n}(Q_n(\theta) - Q_n(\theta_0)) \geq -C_1 n^{1/2-2\psi}\right) \\
&\leq P\left(\sqrt{n}\mathbb{P}_n \tilde{\ell}_0 \gtrsim (C_1/2) n^{1/2-2\psi} + O_P(n^{1/2-3\psi})\right) \\
&\quad + P\left(\sup_{\mathcal{N}_n^c} \sqrt{n}(Q_n(\theta) - Q_n(\theta_0)) \geq -C_1 n^{1/2-2\psi}\right) \\
&\quad + E\left[P\left(\min_{\mathcal{S}_n \cap \mathcal{N}_n} ((\sqrt{n}/2)(\theta - \theta_0)^T \tilde{I}_0 (\theta - \theta_0)) > (C_1/2) n^{1/2-2\psi} | \mathcal{S}_n\right)\right]
\end{aligned}$$

The first two quantity in the last inequality approaches to zero by choosing proper C_1 and C according to similar analysis in the proof of theorem 1. We

next analyze the last quantity.

$$
\begin{aligned}
& E\left[P\left(\min_{\mathcal{S}_n\cap\mathcal{N}_n}((\sqrt{n}/2)(\theta-\theta_0)^T\tilde{I}_0(\theta-\theta_0))>(C_1/2)n^{1/2-2\psi}|\mathcal{S}_n\right)\right] \\
& \leq E\left[P\left(\min_{\mathcal{S}_n}((\theta-\theta_0)^T\tilde{I}_0(\theta-\theta_0))>C_1n^{-2\psi}|\mathcal{S}_n\right)\right] \\
& \leq \left[1-P\left(\|\bar{\theta}-\theta_0\|^2\lesssim C_1c/card(S_n)\right)\right]^{card(S_n)} \\
& \leq (1-\rho C_1/card(S_n))^{card(S_n)}\to\exp(-\rho C_1),
\end{aligned}
$$

where $\rho>0$. The second inequality follows since the cardinality of $\mathcal{S}_n$ is larger than $cn^{2\psi}$. The last inequality follows from the boundedness of the density of $\bar{\theta}$ around θ_0. This completes the proof of corollary 1. □

Proof of theorem 2: We first prove the below lemma 2.1.

lemma 2.1. Assuming the conditions in theorem 2 and that

$$
\Pi_n(\hat{\theta}_n^{(k-1)},t_n)-\tilde{I}_0=O_P(r_n^{(k-1)}), \tag{6.6}
$$

we have

$$
\begin{aligned}
(\hat{\theta}_n^{(k)}-\hat{\theta}_n) \;=\; O_P\Bigg(&\|\hat{\theta}_n^{(k-1)}-\hat{\theta}_n\|r_n^{(k-1)}\vee\frac{g_r(|s_n|\vee n^{-1/2}\vee\|\hat{\theta}_n^{(k-1)}-\hat{\theta}_n\|)}{n|s_n|} \\
&\vee|s_n|\Bigg)
\end{aligned} \tag{6.7}
$$

for $k=1,2,\ldots$.

Proof: Based on (1.4), we have

$$
\begin{aligned}
\Pi_n(\hat{\theta}_n^{(k-1)},t_n)\sqrt{n}(\hat{\theta}_n^{(k)}-\hat{\theta}_n) \;&=\; \left[\sqrt{n}\Pi_n(\hat{\theta}_n^{(k-1)},t_n)(\hat{\theta}_n^{(k-1)}-\hat{\theta}_n)\right]+\sqrt{n}\Gamma_n(\hat{\theta}_n,s_n) \\
&+\; \left[\sqrt{n}(\Gamma_n(\hat{\theta}_n^{(k-1)},s_n)-\Gamma_n(\hat{\theta}_n,s_n))\right].
\end{aligned} \tag{6.8}
$$

The second term in the above equation equal to

$$
O_P\left(\sqrt{n}|s_n|\vee\frac{g_r(|s_n|)}{\sqrt{n}|s_n|}\right)
$$

according to (3.6) and (6.4). The third term in (6.8) can be written as

$$-\sqrt{n}\tilde{I}_0(\hat{\theta}_n^{(k-1)} - \hat{\theta}_n) + O_P\left(\sqrt{n}|s_n| \vee \frac{g_r(n^{-1/2} \vee |s_n| \vee \|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|)}{\sqrt{n}|s_n|}\right).$$

for $k = 1, 2, \ldots$ by replacing $\hat{\theta}_n^{(0)}$ with $\hat{\theta}_n$ and U_n with $\hat{\theta}_n^{(k-1)} - \hat{\theta}_n$ in (6.2). Combining the above analysis, the assumption (6.6) and nonsingularity of $\tilde{I}_0$, we complete the proof of (6.7). □

We next start the proof of (3.8)-(3.10). Combining (6.3) with (6.7), we can obtain that

$$\begin{aligned}\hat{\theta}_n^{(k)} - \hat{\theta}_n &= O_P\left(\frac{g_r(|t_n| \vee \|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|)\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|}{nt_n^2}\right. \\ &\qquad \left. \vee |s_n| \vee \frac{g_r(|s_n| \vee n^{-1/2} \vee \|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|)}{n|s_n|}\right) \\ &= O_P(f_{k-1}(|t_n|) \vee g_{k-1}(|s_n|)).\end{aligned}$$

Considering the form of $g_r(\cdot)$ specified in lemma 1, we have $g_{k-1}(|s_n|) \geq \tilde{g}(|s_n|) \equiv (|s| \vee n^{-2r-1/2}|s_n|^{-1} \vee n^{-3/2}|s_n|^{-1})$. The smallest convergence rate for $\tilde{g}(|s_n|)$ is $n^{-3/4}(n^{-r-1/4})$ if we choose $s_n \asymp n^{-3/4}$ ($s_n \asymp n^{-r-1/4}$). The above analysis implies (3.8).

In the below proof of (3.9) and (3.10), we consider different cases when $r \geq 1/2$ and $1/4 < r < 1/2$, respectively. For $r \geq 1/2$, by some algebra we can show that for $k \geq 1$

$$\hat{\theta}_n^{(k)} - \hat{\theta}_n = O_P(\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^{3/2}) \tag{6.9}$$

when $\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^{-1} = O_P(\sqrt{n})$, $s_n^* \asymp \|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^{3/2}$ and $t_n^* \asymp \|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|$. And when $\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\| = O_P(n^{-1/2})$, $\|\hat{\theta}_n^{(k)} - \hat{\theta}_n\|$ achieves the optimal rate $O_P(n^{-3/4})$ given $s_n^* \asymp n^{-3/4}$ and $t_n^* \asymp n^{-1/2}$. Thus we only need to figure out how many iterative steps needed for k-step MLE to achieve root-n rate. From (6.9), we know that the convergence rate for N_1-step MLE will be $O_P(n^{-1/2})$, where $N_1 = int[\log 2\psi / \log(2/3)]$, given $\hat{\theta}_n^{(0)}$ is n^{ψ}-consistent. This concludes the proof for $r \geq 1/2$.

We next show (3.10) when $1/4 < r < 1/2$. Similarly we have for $k \geq 1$

$$\hat{\theta}_n^{(k)} - \hat{\theta}_n = O_P(\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^{3/2}) \tag{6.10}$$

if $\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^{-1} = O_P(n^r)$, $s_n^* \asymp \|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^{3/2}$ and $t_n^* \asymp \|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|$. However if $\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^{-1} = O_P(n^{1/2})$ and $\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\| = O_P(n^{-r})$, we have for $k \geq 1$

$$\hat{\theta}_n^{(k)} - \hat{\theta}_n = O_P(\|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^{1/2} n^{-r}) \tag{6.11}$$

given $s_n^* \asymp \|\hat{\theta}_n^{(k-1)} - \hat{\theta}_n\|^{1/2} n^{-r}$ and $t_n^* \asymp n^{-r}$. We next consider two-stage iterations for K-step MLE. If $\hat{\theta}_n^{(0)}$ is n^{ψ}-consistent for $\psi < r$, then at least M_1 iterations are needed such that $\|\hat{\theta}_n^{(M_1)} - \hat{\theta}_n\| = O_P(n^{-r})$ based on (6.10), where $M_1 = int[\log(\psi/r)/\log(2/3)]$. When K-step MLE has achieved the n^r-consistency, we further need M_2 steps to achieve the root-n rate, where $M_2 = int[\log(4r/(4r-1))/\log(2) - 1]$, from (6.11). Then we complete the whole proof for theorem 2. □

References

Bickel, P. J., Klaassen, C. A. J., Ritov, Y. and Wellner, J. A. (1998). *Efficient and Adaptive Estimation for Semiparametric Models*. Springer-Verlag, New York.

Cheng, G. and Kosorok, M.R. (1987) (2006). Higher order semiparametric frequentist inference with the profile sampler. *Annals of Statistics*, Accepted.

Cheng, G. and Kosorok, M.R. (2007). General Frequentist Properties of the Posterior Profile Distribution. *Annals of Statistics*, Invited Revision.

(http://arxiv.org/abs/math.ST/0612191)

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society, Series B* **34** 187–220.

Fan, J. and Chen, J. (1999). One-Step Local Quasi-Likelihood Estimation. *Journal of the Royal Statistical Society, Series B* **61** 927–943.

Fan, J., Lin, H. and Zhou, Y. (2006). Local Partial-Likelihood Estimation for Lifetime Data. *Annals of Statistics* **34** 290–325.

Groeneboom, P. (1991). Nonparametric maximum likelihood estimators for interval censoring and deconvolution. *Technical Report* **378**, Department of Statistics, Stanford University.

Huang, J. (1996). Efficient estimation for the Cox model with interval censoring. *Annals of Statistics* **24,** 540–568.

Jassen, P., Jureckova, J. and Veraverbeke, N. (1985). Rate of Convergence of One- and Two-stpe M-estimators with Applications to Maximum Likelihood and Pitman Esitmators. *Annals of Statistics* **25** 1471–1509.

Kosorok, M. R., Lee, B. L. and Fine, J. P. (2004). Robust inference for univariate proportional hazards frailty regression models. *Annals of Statistics* **32** 1448–1491.

Kosorok, M. R. (2007). *Introduction to Empirical Processes and Semiparametric Inference*. Springer, New York.

Lee, B. L., Kosorok, M. R. and Fine, J. P. (2005). The profile sampler. *Journal of the American Statistical Association* **100** 960–969.

Murphy, S. A. and Van der Vaart, A. W. (1999). Observed information in semiparametric models. *Bernoulli* **5** 381–412.

Murphy, S. A. and Van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association* **93** 1461–1474.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.

Cai, Z., Fan, J. and Li, R. (2000). Efficient Estimation and Inferences for Varying-Coefficient Models. *Journal of the American Statistical Association* **95** 888–902.

Department of Statistical Science
Duke University
Durham, NC, U.S.
E-mail: chengg@stat.duke.edu
Phone: 919-684-5956
Fax: 919-684-8594

Table 1. *Cox regression with right censored data($\theta_0 = 1$ and 500 samples).*

n	$\hat{\theta}_n^{(0)}$	$\hat{\theta}_n^{(1)}$	$\hat{\theta}_n$	$n^{3/4}\lvert\hat{\theta}_n - \hat{\theta}_n^{(1)}\rvert$
50	1.0229	1.0222	1.0167	0.1030
100	1.0346	1.0344	1.0324	0.0632
200	0.9979	0.9979	1.0028	0.2606
500	0.9974	0.9974	0.9964	0.1057

Table 2. *Cox regression with current status data ($\theta_0 = 1$ and 500 samples).*

n	$\hat{\theta}_n^{(0)}$	$\hat{\theta}_n^{(1)}$	$\hat{\theta}_n^{(2)}$	$\hat{\theta}_n^{(3)}$	$\hat{\theta}_n$	$n^{7/12}\lvert\hat{\theta}_n - \hat{\theta}_n^{(3)}\rvert$
50	1.0452	1.8218	1.7226	1.7563	1.1962	5.4870
100	0.8017	0.7604	0.7997	0.8289	0.8541	0.3699
200	0.8118	0.7692	0.8474	0.8425	0.8859	0.9545
500	0.8376	0.8364	0.8757	0.9592	0.9896	1.1410

n, sample size; $\hat{\theta}_n^{(0)}$, initial estimate; $\hat{\theta}_n^{(1)}$, one step MLE; $\hat{\theta}_n^{(2)}$, two step MLE; $\hat{\theta}_n^{(3)}$, three step MLE; $\hat{\theta}_n$, MLE.